Provably Efficient RL under Episode-Wise Safety in Constrained MDPs with Linear Function Approximation

Toshinori Kitamura, Arnob Ghosh, Tadashi Kozuno, Wataru Kumagai,

Kazumi Kasaura, Kenta Hoshino, Yohei Hosoe, Yutaka Matsuo

Contribution

We develop the first linear CMDP algorithm that guarantees:

- episode-wise safe exploration
- $\widetilde{\mathcal{O}}(\sqrt{K})$ regret
- and computational tractability

Table 1: $\widetilde{\mathcal{O}}(\sqrt{K})$ regret CMDP algorithms

	paper	Epi.-wise safe?	Comp. Efficient?
Tabular	Yu et al. [47]	Yes	State size dependent
Linear	Ghosh et al. [18]	No	No
	Roknilamouki et al. [36]	Instantaneous	Yes
	Ours	Yes	Yes

Notation: Linear CMDP

Linear CMDP: $(\mathcal{S}, \mathcal{A}, H, P, r, u, b, s_1)$

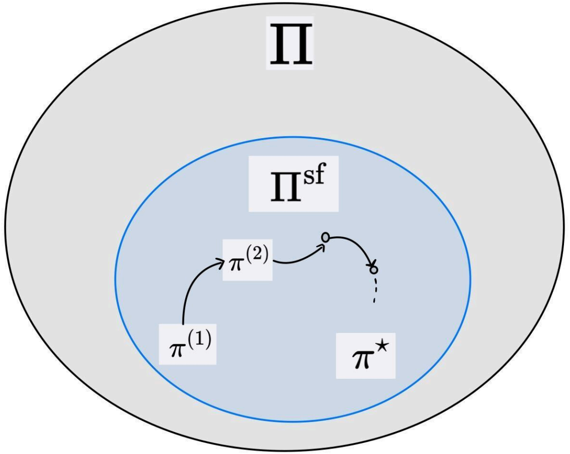
- Finite state & action spaces: $\mathcal{S}$, $\mathcal{A}$
- Horizon: $H \in \mathbb{N}$
- Transition probability: $P_h(s' \mid s, a) \in [0, 1]$
- Reward & utility: $r_h(s, a), u_h(s, a) \in [0, 1]$
- Constraint threshold: $b \in [0, H]$
- Initial state: $s_1 \in \mathcal{S}$

Policy and Value Function

- Value functions: $V^{\pi,r}_{P,h}, V^{\pi,u}_{P,h} : \mathcal{S} \to \mathbb{R}$, where $V^{\pi,r}_{P,h}(s) = \mathbb{E}\left[\sum_{h'=h}^{H} r_{h'}(s_{h'}, a_{h'}) \mid s_h = s\right]$
- Action value functions: $Q^{\pi,r}_{P,h}, Q^{\pi,u}_{P,h} : \mathcal{S} \to \mathbb{R}$, where $Q^{\pi,r}_{P,h}(s, a) = \mathbb{E}\left[\sum_{h'=h}^{H} r_{h'}(s_{h'}, a_{h'}) \mid s_h = s, a_h = a\right]$

- $P_h(s' \mid s, a) = \boldsymbol{\mu}_h(s')^\top \boldsymbol{\phi}(s, a)$
  - Known feature map $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$
  - Unknown vectors $\boldsymbol{\mu}_h := \left(\boldsymbol{\mu}_h^1, \ldots, \boldsymbol{\mu}_h^d\right) \in \mathbb{R}^{S \times d}$
- $r_h(s, a) = (\boldsymbol{\theta}_h^r)^\top \boldsymbol{\phi}(s, a)$, $u_h(s, a) = (\boldsymbol{\theta}_h^u)^\top \boldsymbol{\phi}(s, a)$
  - Known vectors $\boldsymbol{\theta}_r, \boldsymbol{\theta}_u \in \mathbb{R}^d$

Setting: Episode-wise Safe Exploration



- Policies $\pi^{(1)}, \dots, \pi^{(K)} \in \Pi$
- Safe policy set: $\Pi^{\text{sf}} \triangleq \left\{\pi \mid V_{P,1}^{\pi,u}(s_1) \geq b\right\}$
  - Episode-wise safety: $\pi^{(k)} \in \Pi^{\text{sf}}$ for all k

Assumption (Slater condition): We have access to $\pi^{\text{sf}} \in \Pi^{\text{sf}}$ and $\xi > 0$ such that $V_{P,1}^{\pi^{\text{sf}},u}(s_1) \geq b + \xi$.

Goal: Sublinear regret with episode-wise safety

$$\text{Regret}(K) := \sum_{k=1}^{K} V_{P,1}^{\pi^{\star},r}(s_1) - V_{P,1}^{\pi^{(k)},r}(s_1) = o(K) \ \text{ such that } \ \pi^{(k)} \in \Pi^{\text{sf}} \quad \forall k \in [\![1, K]\!]$$

where $\pi^{\star} \in \arg\max_{\pi \in \Pi^{\text{sf}}} V_{P,1}^{\pi,r}(s_1)$ is an optimal safe policy.

Basic Approach: Optimistic-Pessimistic (Opt-Pes) Exploration

Idea (e.g., Yu et al. [47]): Estimate reward value optimistically and utility value pessimistically:

$$\text{(Opt-Pes)} \quad \pi^{(k)} \in \arg\max_{\pi \in \Pi} \underbrace{\overline{V}^{\pi,r}_{(k),1}(s_1)}_{\geq V^{\pi,r}_{P,1}(s_1)} \text{ such that } \underbrace{\underline{V}^{\pi,u}_{(k),1}(s_1)}_{\leq V^{\pi,u}_{P,1}(s_1)} \geq b\,. \tag{1}$$

$\overline{V}^{\pi,r}_{(k),1} := \widehat{V}^{\pi,r+\beta^{(k)}}_{(k),1}$ where $\widehat{V}^{\pi,r}_{(k),1}$ is the value estimate and $\beta^{(k)}$ is the bonus. Similarly, $\underline{V}^{\pi,u}_{(k),1} := \widehat{V}^{\pi,u-\beta^{(k)}}_{(k),1}$.

- ☺️ Optimism leads to sublinear regret
- ☺️ Pessimism ensures $\pi^{(k)} \in \Pi^{\mathrm{sf}}$

Technical contributions

We introduce two main techniques to address these challenges in linear CMDPs:

1. Tractability — The large state space makes optimizing Eq. (1) non-trivial.
2. Feasibility — Pessimistic estimates can make Eq. (1) infeasible.

Technique 1: Efficient Implementation of Opt-Pes Policy

$$\text{(Opt-Pes)} \quad \pi^{(k)} \in \arg\max_{\pi\in\Pi} \overline{V}^{\pi,r}_{(k),1}(s_1) \text{ such that } \underline{V}^{\pi,u}_{(k),1}(s_1) \geq b\,, \tag{1}$$

Instead of solving Opt-Pes problem, we realize Opt-Pes by "softmax policy" using estimated value functions:

$$\pi_h^{(k),\lambda}(\cdot \mid s) = \mathrm{SoftMax}\left(\overline{Q}^{r}_{(k),h}(s,\cdot) + \lambda \underline{Q}^{u}_{(k),h}(s,\cdot)\right)$$

Lemma 6:

- If $\lambda \approx 0$, then $\pi^{(k),\lambda}$ favors optimistic exploration
- If $\lambda \gg 0$, then $\pi^{(k),\lambda}$ tries to satisfy the pessimistic constraint

$$\overline{\underline{Q}}^{r \text{ or } u}_{(k),1} := \widehat{Q}^{\pi, r\pm\beta^{(k)}}_{(k),1}$$

We do bisection search in the interval $\lambda \in [0, C_\lambda]$ to find a good λ:

- If $\underline{V}^{\pi^{(k),\lambda},u}_{(k),1}(s_1) \geq b$, then $\pi^{(k),\lambda}$ is safe. Decrease $\lambda \downarrow$
- Otherwise, $\pi^{(k),\lambda}$ is unsafe. Increase $\lambda \uparrow$.

$C_\lambda > 0$ is an upper bound of λ obtained by the Slater condition.

Technique 2: Safe Policy Deployment

$$\text{(Opt-Pes)} \quad \pi^{(k)} \in \arg\max_{\pi \in \Pi} \overline{V}^{\pi,r}_{(k),1}(s_1) \ \text{ such that } \ \underline{V}^{\pi,u}_{(k),1}(s_1) \geq b \, . \tag{1}$$

At the start of training, the bonus may be so large that no policy satisfies the pessimistic constraint.

Safe policy deployment technique: Deploy π^{sf} if the policy $\pi^{(k),C_\lambda}$ is unsafe: $\underline{V}^{\pi^{(k),C_\lambda},u}_{(k),1}(s_1) < b$

$\rightarrow$ For any k, $\pi^{(k)}$ is ensured to be safe: it is chosen by either

1. π^{sf} that is safe or 2. $\pi^{(k),\lambda}$ that satisfies the pessimistic constraint.

🤔 For sublinear regret, we need to bound the # of π^{sf} deployments...

Theorem 3. With high probability, the safe policy π^{sf} is deployed at most $\widetilde{\mathcal{O}}(d^3 H^4 \xi^{-2})$ times.

(See Section 2 for an intuitive derivation.)

Algorithm : Optimistic-Pessimistic Softmax Exploration for Linear CMDP

For each $k = 1, 2, \ldots, K$, do

1. If $\underline{V}^{\pi^{(k),C_\lambda},u}_{(k),1}(s_1) < b$, all the softmax policies may not be safe. Deploy $\pi^{(k)} = \pi^{\mathrm{sf}}$.
2. Else, find the best λ for the softmax policy:
   - If $\underline{V}^{\pi^{(k),\lambda},u}_{(k),1}(s_1) \geq b$, λ is large. Decrease $\lambda \downarrow$.
   - Else, increase $\lambda \uparrow$.
   - After sufficient iterations, set $\pi^{(k)} = \pi^{(k),\lambda}$.
3. Sample a trajectory $\left(s_1^{(k)}, a_1^{(k)}, \ldots, s_H^{(k)}, a_H^{(k)}\right)$ by deploying $\pi^{(k)}$.

Main Open Problem: Can we achieve episode-wise safety in linear CMDPs with multiple constraints?

Theorem 4 With high probability, the algorithm achieves

$$\pi^{(k)} \in \Pi^{\mathrm{sf}} \forall k \in \{1, \ldots, K\} \quad \text{and} \quad \mathrm{Regret}(K) \leq \widetilde{\mathcal{O}}\left(H^4 \xi^{-1} \sqrt{d^5 K}\right)$$